LA DISTRIBUCIÓN DEL TAMAÑO DE LAS CIUDADES
La Ley de Zipf revisitada

Josep Roca[1] y Blanca Arellano[2]

[1] Doctor Arquitecto, Director del Centro de Política del Suelo y Valoraciones, Universitat Politècnica de Catalunya, Barcelona España, josep.roca@upc.edu
[2] Maestra Arquitecta, Jefa de Vivienda en SIDUE y Profesora de Estudios Urbanos IV, Universidad Autónoma de Baja California, Mexicali México, blanca.arellano@uabc.edu.mx

Resumen

Diversos autores (Berry 1970, Krugman 1996 o Eaton y Eckstein 1997, entre muchos otros) han experimentado asombro acerca de cómo en la mayor parte de lugares se cumple con una claridad meridiana la ley del "mínimo esfuerzo" establecida por Zipf (1949). Las ciudades, ordenadas por población, parecen seguir casi al pie de la letra una función log/log, en la que el logaritmo de la "masa" (población, densidad, número de trabajadores, etc.) se correlaciona de forma casi perfecta con el logaritmo del orden de dicha masa. Esa función log/log, avanzada ya por Pareto en el siglo XIX, ha seducido a no pocos investigadores, al producirse, en hipótesis, tanto en fenómenos naturales (terremotos, meteoritos, especies vivas, …) como derivados de la sociedad (lenguaje, o distribución de ciudades), lo que ha llevado a indagar en sus fundamentos teóricos (Simon 1955, Brakman et al. 1999, Gabaix 1999).

Si bien algunos autores (Rosen y Resnick 1980, Fan y Casetti 1994) han discutido la validez lineal de la Ley Zipf, introduciendo modelos no lineales, la literatura especializada se ha concentrado en la "cola superior" de la jerarquía urbana, las ciudades o áreas metropolitanas grandes, tendiendo a callar el hecho de que la función log/log en absoluto parece ser un modelo de alcance general. La ponencia intenta mostrar que cuando se tiene en cuenta la totalidad de casos (es decir, la totalidad de localidades pobladas en un determinado territorio), el modelo log/log parece ser tan sólo un caso singular propio de "los grandes". De hecho se pone en evidencia que un modelo log/lin tiende a ser más eficiente, aunque se "dobla por las colas". Ello ha conducido a la hipótesis que intenta ser contrastada en esta investigación que el logaritmo de la masa urbana tiende a tener una "distribución normal", conduciendo su distribución acumulada (y ordenada por rango) a distribuirse de acuerdo con una estructura de carácter logístico, en "S".

En este sentido la observación reiterada del cumplimiento de la Ley de Zipf en el tamaño de las ciudades sería tan sólo la punta emergida de un "iceberg" más profundo, en el que ciudades medias y pequeñas tienen también su protagonismo, y donde una "ley" de carácter más general emerge.

La investigación presentada se interroga acerca de si puede modelarse de forma sencilla y elegante esa emergencia "normal" del logaritmo de la masa e intenta algún experimento en este sentido.

1.- Introducción

Una de las cuestiones que más ha atraído a los especialistas urbanos, en especial a los economistas, ha consistido en la comprensión de la jerarquía espacial inherente a la distribución de los tamaños de las ciudades. Desde hace décadas es bien conocido que la distribución de las grandes ciudades en muchos lugares del mundo puede ser descrita por una ley de tipo exponencial, descrita en su forma más popular por Zipf (1949)[1], según la cual el número de ciudades con una población mayor que "*P*" es aproximadamente proporcional a P^{-a}, estando "*a*" muy próximo a 1. Zipf, estudiando la frecuencia de uso de las palabras de lengua inglesa, en función a su complejidad (el número de letras), llegó a formular una "ley", que denominó de "mínimo esfuerzo", según la cual la frecuencia de una palabra, P_n, ordenada en el orden *n* tendría una frecuencia igual a:

$$P_n \sim 1/n^a$$

Dicha "ley" ha sido contrastada para multitud de fenómenos naturales o artificiales. Desde la frecuencia de los terremotos en virtud a su magnitud, hasta al tamaño de las ciudades.

La razón de por qué la distribución del tamaño de las ciudades sigue la "ley" de Zipf ha intrigado a no pocos teóricos (Simon, 1955; Henderson, 1974; Krugman, 1996; Brakman et al., 1999; Gabaix, 1999[2]), lo que ha llevado a Krugman (1999) a decir:

> *"Llegados a este punto, no tenemos en nuestro haber ninguna explicación de la asombrosa regularidad que presentan las distribuciones del tamaño de las ciudades. Debemos reconocer que este hecho supone un verdadero desafío intelectual"*[3].

El origen de la "ley" descrita por Zipf parece encontrarse en el estudio de la distribución de rentas llevado a cabo por Pareto (1896), y según la cual se produciría el conocido efecto "80-20"[4]. Ya en 1913, Auerbach (1913) propuso que la distribución del tamaño de las ciudades podría estar muy cercana a una distribución de Pareto. Entonces si ordenamos las ciudades desde la más grande

---

[1] Ya en 1682, Alexandre Le Maître reconoció la existencia, en Francia, de una clara estructura en la distribución del tamaño de las ciudades. Pero no fue sino hasta 1913 que Felix Auerbach que se estableció formalmente la estructura de la relación matemática, que más tarde Zipf generalizaría en la función exponencial con exponente -1.

[2] Desde Simon (1955) se ha venido a discutir la vinculación entre la distribución de Pareto con el principio desarrollado por Gibrat (1931), según el cual no existe relación entre la tasa de crecimiento (de las ciudades en nuestro caso) y la dimensión inicial, razón por la cual no se puede deducir comportamiento regular alguno. Véase, entre otros, Gabaix (1999).

[3] Véase Fujita, M., Krugman, P. & Venables, A.J. (1999), pág 223 de la edición española de 2000.

[4] Es decir, que el 20% de la población acumularía el 80% de la riqueza.

(rango 1) a la más pequeña (rango N), el rango de una ciudad de población P, r(P), sería:

$$r(P) = AP^{-\alpha}$$

En logaritmos:

$$Ln\ r(P) = Ln\ A - \alpha\ Ln\ P$$

Siendo $\alpha = 1$, un caso particular de la distribución de Pareto, en la interpretación dada por Zipf.

El trabajo empírico realizado a lo largo de varias décadas (Berry, 1961; Berry & Horton, 1970; Rosen & Resnick, 1980; Carroll, 1982; Guérin-Pac, 1995; Eaton & Eckstein, 1997; Chesire, 1999; Dobkins & Ioannides, 2000) parece llegar a la conclusión de que, *en las grandes ciudades*, la distribución de las ciudades sigue, por regla general, la distribución de Pareto. Respecto a si $\alpha = 1$, algunos autores, especialmente Krugman (1996), han defendido con ardor la validez de la tesis de Zipf. Mientras que otros, como Alperovich (1993), la han rechazado.

También en España, Lasuén (1967), primero, y Lanaspa *et alt.* (2004), más recientemente, han confirmado la validez de la distribución exponencial de Pareto. Según estos últimos, que han estudiado la serie temporal de la población de los municipios más grandes españoles desde 1900 a 1999, los grados de ajuste de los modelos logarítmicos desarrollados son óptimos, con niveles de explicación (R^2) superiores al 0,98, siendo siempre el *exponente de Pareto* significativo estadísticamente. Respecto a la "ley" de Zipf su trabajo concluye que el parámetro α es, en todos los modelos ensayados, estadísticamente diferente a uno, con lo que afirman que, para el caso español, no existe evidencia a favor de la citada "ley".

En los últimos años la mayor parte de la literatura, aceptado el principio de Pareto, ha ido dirigida al análisis de *la forma de esa distribución*. Dobkins & Ioannides (2000) han encontrado de que el coeficiente α ha disminuido a lo largo del siglo XX para las ciudades de USA. Resultados similares a los obtenidos por Lanaspa *et alt.* (2004), los cuales encuentran caídas regulares del *coeficiente de Pareto* desde 1900 hasta 1970, así como incrementos en la citada pendiente a partir de esa fecha, lo que es interpretado como una demostración de los cambios operados en la estructura urbana española[5]. Siguiendo a Suárez-Villa (1988), dichos autores

[5] Para los autores citados, "la estructura urbana española experimenta un profundo cambio en su evolución alrededor de mediados de los años setenta. Hasta esa fecha la distribución es cada vez menos igualitaria, de forma que se acentúan las diferencias entre los tamaños de las ciudades, siendo éstas mayores en la parte alta (ciudades más grandes) de la distribución. (…) A mediados de los setenta, y hasta 1999, el panorama se altera y la concentración de la población en los mayores núcleos llega a su tope. La distribución de los tamaños de las ciudades se vuelve en su

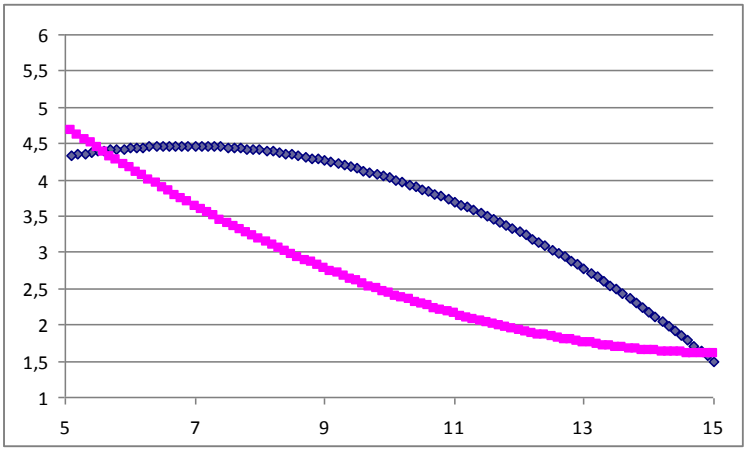
interpretan la evolución del coeficiente de Pareto como *índice de metropolitanización*, confirmando para el caso español la hipótesis de Parr (1985) acerca de la evolución en U de dicho exponente a lo largo del tiempo en los países desarrollados.

En el análisis de la forma de distribución del tamaño/rango de las ciudades, algunos autores han propuesto transformaciones respecto al modelo clásico de Pareto. La literatura especializada (Lanaspa *et alt.* 2004) ha destacado que si bien las distribuciones paretianas se ajustan razonablemente bien a la distribución del tamaño de las ciudades, se puede plantear de forma complementaria la posibilidad de que la relación entre rango y tamaño no sea de carácter lineal (Rosen y Resnick, 1980; Fan y Casetti, 1994). Particularmente ha tenido amplia difusión la transformación cuadrática:

$$Ln\ r(P) = Ln\ A - \alpha\ Ln\ P + \beta\ Ln\ P^2$$

Donde β permitiría comprender la constatada curvatura de las "colas" de la distribución rango/tamaño. El planteamiento teórico sugiere, siguiendo a Pareto, que el coeficiente α debiera ser positivo, ponderando el β el grado de metropolitanización de la estructura del sistema urbano: creciente (es decir con elevada "macrocefalia"), en caso de ser de signo positivo (color lila en la figura n. 1) y decreciente (en estructuras tendentes a una mayor dispersión), si negativo (azul en la figura)[6].

Figura n. 1: Modelos cuadráticos



El contraste empírico de esta estructura no lineal ha sido confirmada estadísticamente para los principales municipios españoles por Lanaspa *et alt.* (2004), con R^2 sorprendentemente altas (superiores a 0,99), y con un acusado

conjunto menos desigual, de forma que son las aglomeraciones pequeñas y medianas las que ahora crecen más deprisa" (Lanaspa *et alt.* 13-14).

[6] Para Lanaspa *et alt* (2004), si $\beta = 0$ nos encontraríamos con la Ley de Gibrat (1931).

cambio en la curvatura de las colas (*β* pasa de tener signo positivo a negativo, a partir de 1970).

La mayor parte de los trabajos empíricos desarrollados, sin embargo, parecen obedecer a una voluntad implícita de querer demostrar la validez de la "ley" de Zipf, o cuando menos de su versión menos restrictiva de Pareto. La velada advertencia de que la relación log-log es válida sólo para las "ciudades" suele conducir, en el mejor de los casos, a una definición, abstracta a juicio de los autores de este trabajo, de lo que es *ciudad*. Así, por ejemplo, el trabajo de Rosen & Resnick (1980) se centra en las 50 principales ciudades de 44 países. Krugman (1996) se limita a las 130 principales áreas metropolitanas USA. Dobkins & Ioannides (2000) al conjunto de áreas metropolitanas, desconociendo el hecho de las ciudades menores. Lanaspa *et alt.* (2004) a los 100-300-700 municipios mayores de España. Y así casi en toda la literatura. Berry & Horton (1970) se refieren al límite de 250.000 habitantes como "the size of urban region (…) to constitute the minimum threshold scale for economic and social viability in contemporary, metropolitanized America"[7].

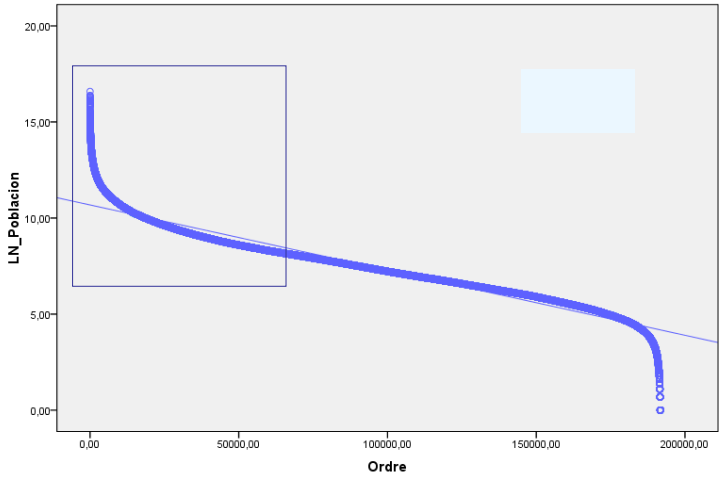
Rosen and Resnick (1980), en su trabajo acerca de la literatura desarrollada entre 1950 y 1970, subrayan la importancia de la definición de ese "lower threshold size for cities", aspecto que también ha sido subrayado, de forma más reciente, por Dobkins & Ioannides (2000) así como por Black & Henderson (2003). Sin embargo, desde 1980, no se ha profundizado de forma significativa acerca de la definición de "ciudad", que estaría detrás de las fuerzas económicas que conducen a la distribución de Pareto. La limitación a umbrales de población, o a conceptos administrativos de *área metropolitana*, no parece el camino a seguir en la investigación.

Este trabajo se centra en la discusión de la validez de la "ley" de Zipf, así como de la distribución de Pareto, cuando se analiza la *totalidad del sistema urbano*, y no tan sólo su "cola" superior. En ese caso emergen formas muy distintas de la relación rango/tamaño, en las que la relación log-log es tan sólo una parte singular de las mismas: limitada al "upper tail" de la distribución completa. La distribución de Pareto, en este caso, se limitaría a explicar la punta iceberg, haciendo necesaria una teoría más completa que se plantee comprender acerca de si existen regularidades en la relación rango/tamaño de las ciudades, así como las causas inherentes a las mismas.

En este sentido en este trabajo se propone, siguiendo a Eeckhout (2004), que el aparente cumplimiento de la distribución de Pareto en los sistemas urbanos de gran magnitud (p.e. > 1.000.000 habitantes) obedece en realidad a la visión sesgada del "upper tail" del sistema territorial completo, el cual, aparentemente, seguiría, en su totalidad, una distribución log-normal.

[7] Berry & Horton, *op cit*, pág. 64.

Figura n. 2: ¿Una singularidad del "upper tail"?



Como ha indicado Eeckhout (2004): "At the very upper tail of the distribution, there is no dramatic difference between the density function of the lognormal and the Pareto. Now both the truncated lognormal and the Pareto density are downward sloping and similar (the Pareto is slightly more convex). As a result, both the Pareto and the truncated lognormal trace the data relatively closely" [8] . Aparentemente podrían producirse simultáneamente, por tanto, ambas leyes: la distribución log-log de Pareto en el "upper tail", como una singularidad parcial del conjunto normal del logaritmo del tamaño. Basado en los datos del U.S. Census 2000 data, Eeckhout (2004) sostiene que el conjunto de la distribución de las ciudades norteamericanas adopta una forma lognormal antes que paretiana, contrastando dicha hipótesis mediante la aplicación del test de Kolmogorov-Smirnov (KS) para distribuciones normales.

Dicha propuesta, criticada por Levy (2009)[9], es actualmente objeto de discusión por parte de la literatura especializada. González-Val *et alt.* (2008) han encontrado evidencias de la distribución lognormal para el conjunto de las unidades urbanas de Italia, España y USA, de 1900 a la actualidad, utilizando para ella una aplicación específica del test de Wilcoxon sobre verificación de la hipótesis nula de igualdad entre distribuciones. Por su parte Malevergne, *et alt.* (2009) confirman la validez estadística de la distribución de Pareto para las primeras 1.000 ciudades

[8] Ecckhout (2004), pág. 1432.

[9] Levy (2009) argumenta que el 0.6% superior de las ciudades norteamericanas, el cual agrupa más del 23% de la población se separa drásticamente de la distribución lognormal, mostrando una mayor congruencia con la hipótesis log-log de Pareto. Para Levy, si bien el grueso de la distribución sigue la ley lognormal, en el upper tail no puede ser confirmada mediante la aplicación de los 2-test convencionales. El no rechace, por Eeckhout (2004), de la hipótesis lognormal proviene, para Levy, del uso del test de Lilienfords (L test), el cual viene dominado por el centro de la distribución, antes que por sus colas, "where the interesting action occurs".

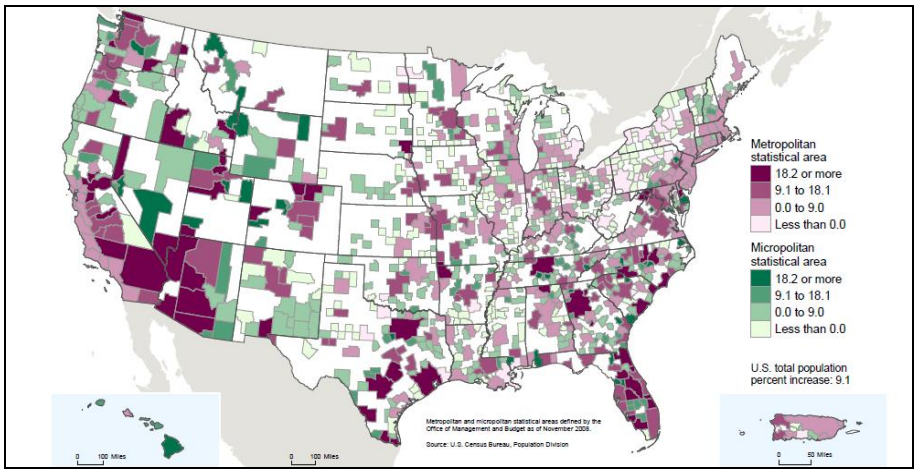
USA, si bien sugieren que para las de menor tamaño la distribución log-normal es más eficiente.

El trabajo que aquí se presenta, busca mostrar evidencias empíricas acerca del citado debate, sugiriendo vías alternativas de desarrollo. Especialmente se sostiene que la ley log-normal aparece de forma más evidente cuando se considera las *ciudades reales* antes que las simples *unidades administrativas*, revelándose como un instrumento eficiente para la comprensión del fenómeno de distribución del tamaño de los sistemas urbanos.

2.- Un primer análisis empírico: las áreas metro y micropolitanas USA.

Primero de todo replicaremos el análisis de la distribución tamaño/rango de los sistemas urbanos USA, sin la restricción de referirnos al upper tail de las áreas metropolitanas (Metro). A tal fin incluiremos no sólo las Metro sino también las áreas micropolitanas (Micro)[10], tal como las definió el Census Bureau para 2000. Eso nos permite trabajar no sólo con los sistemas de más de 100.000 habitantes (385 Metropolitan Areas), sino con los 940 sistemas urbanos que superan los 10.000 habitantes, según datos de 2009.

Figura n. 2: Evolución de la población de las áreas micro y metropolitanas (2000-2009)



Fuente: US Census Bureau

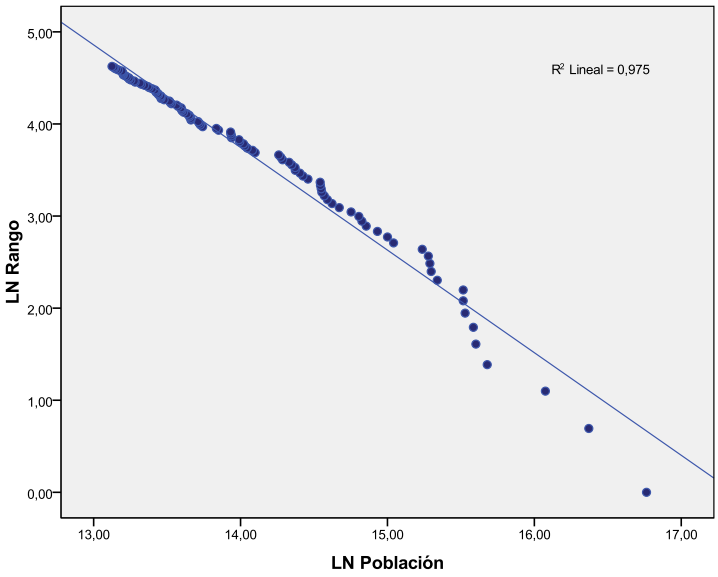
La figura n. 3 nos muestra el resultado de aplicar la distribución de Pareto para los sistemas urbanos de más de 500.000 habitantes (102 Metro). Como se puede observar la relación log-log parece confirmarse (R^2 = 0,975), aunque no tanto la

[10] El Censo USA diferencia las áreas metropolitanas de las micropolitanas. Las primeras con un county o ciudad central de 50.000 o más habitantes, que agrega un sistema urbano de (en su conjunto) más de 75.000 habitantes. El área metropolitana se consolida a partir de los counties (o ciudades) que envían más de un determinado porcentaje de sus residentes a trabajar al corazón del sistema urbano. Por su parte las áreas micropolitanas se delimitan siguiendo u procedimiento similar, si bien el centro urbano puede alcanzar un umbral mínimo de 10.000 habitantes.

"ley" de Zipf, ya que el exponente de Pareto es estadísticamente distinto a 1 (-1,114).

En apariencia el modelo log-log continúa funcionando bien (R^2 = 0,974), cuando se considera el conjunto de sistemas urbanos de > 10.000 habitantes, aún cuando la "ley" de Zipf sigue sin ser confirmada (α = - 0,795). Sin embargo la figura n. 4 evidencia la clara concavidad inferior de la distribución, confirmada por el modelo cuadrático ensayado (R^2 = 0,997; α = 0,913; β = - 0,070)[11]. Dicho modelo permite poner en duda la validez de la distribución de Pareto no tanto porque sea significadamente más eficiente que la log-log, sino por el evidente cambio de signo experimentado por el coeficiente α[12]. Dicho cambio se debe a una razón más profunda que la colinearidad existente entre el logaritmo de la población y el cuadrado de ese logaritmo. El cambio se debe a que la verdadera relación que subyace a la muestra estudiada es no tanto la relación log-log, sino log-log^2. El logaritmo de la población explica los residuos no explicados por el logaritmo al cuadrado, y no al revés, como sería de esperar si la distribución de Pareto fuese cierta.

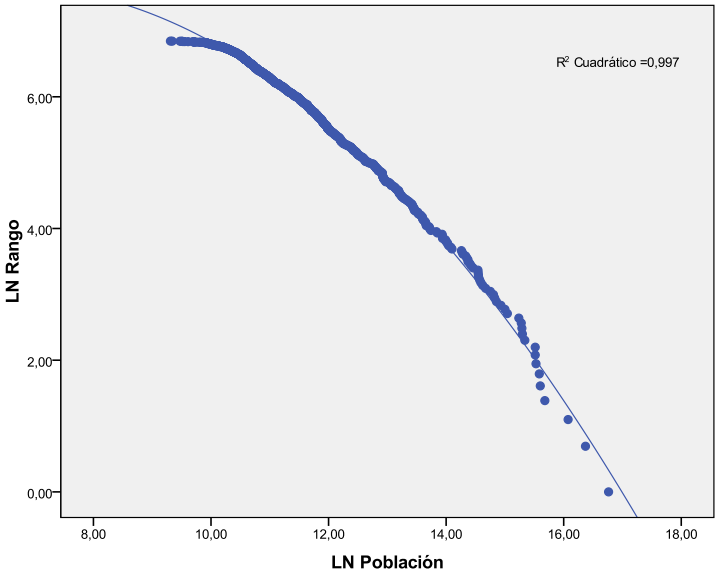
Figura n. 3: Áreas Metropolitanas USA (> 500.000 habitantes, 2009)



[11] Otros trabajos, con datos anteriores, habían dado para las Áreas Metropolitanas USA resultados similares. Así Rosen and Resnick (1980) encontraron, para datos de 1980, esa misma downward concavity.

[12] Idéntico resultado cóncavo inferior se obtiene si la muestra estudiada se limita a las Áreas Metropolitanas de > 500.000 habitantes.

Figura n. 4: Sistemas Micro y Metropolitanos USA (>10.000 habitantes, 2009)



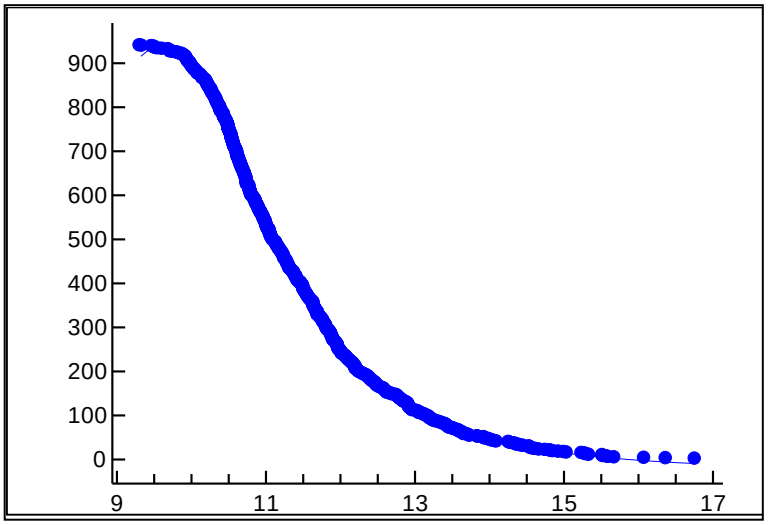
La nada intuitiva interpretación de la ecuación cuadrática del logaritmo, parece sugerir que algo oculto permanece en la distribución rango/tamaño de las ciudades, más allá de la interpretación dada por Pareto.

Lo anterior conduce a analizar la relación rango/tamaño desde una perspectiva no paretiana, del tipo lin/log. En ese caso se hace evidente una distribución en S, que viene contrastada con el buen rendimiento de los modelos de tipo sigmoideo, como el representado en la figura n. 5 (R^2 = 0,999).

Figura n. 5: Distribución de "Racional"[13]

Forma en *S*, que recuerda de forma acusada la distribución acumulada de una distribución normal (cdf en la literatura especializada). Lo anterior conduce a las siguientes preguntas:

- ¿Nos encontramos, simplemente, ante una distribución normal del logaritmo de la población?



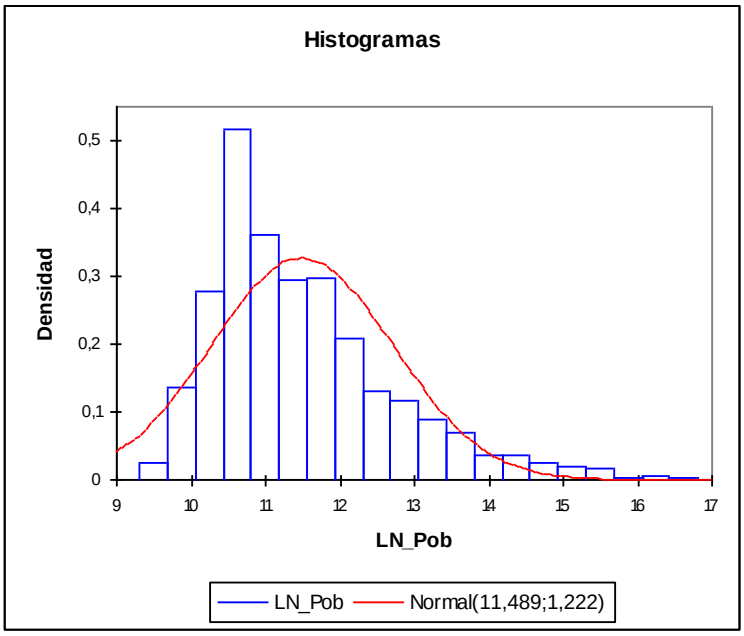
- ¿La distribución de Zipf/Pareto no sería, entonces, simplemente el "upper tail" de dicha distribución normal acumulada?

El test de Kolmogorov-Smirnov (KS) no permite confirmar la hipótesis nula, con lo que parece confirmarse que la distribución del logaritmo de las Áreas Metro y Micropolitanas se aleja de la normalidad[14]. Por otra parte, el análisis del histograma adjunto tampoco permite asegurar la estructura normal del logaritmo de la población. A pesar de que la tendencia del histograma claramente sugiere una estructura normal, la mayor frecuencia de las ciudades situadas en el rango del población (logaritmo) situado entre 10 y 11, respecto al rango "central", entre 11 y 12, permite dudar que realmente la distribución del logaritmo del tamaño obedezca a una ley normal.

[13] La distribución "Racional" ensayada puede ser expresada del modo que sigue:
*R(P) = ((A+(B*x))/((1+(C*x))+(D*(x^2))))*, siendo A, B, C y D constantes.

[14] El test devuelve un valor D = 0,097, lo que corresponde a un p-value < 0,0001, por lo que la hipótesis nula de igualdad con la distribución normal debe ser rechazada.

Figura n. 6: Histograma del LN de la Población



Sin embargo el test KS, como ha destacado la doctrina, tiende a devolver resultados negativos en muestras relativamente grandes, lo que conduce a conducido a la búsqueda de recursos alternativos, como por ejemplo el test de Wilcoxon (W), sugerido por Lanaspa *et alt* (2004), para demostrar la igualdad entre dos distribuciones).

La aplicación del test W implica contrastar la distribución de los rangos obtenidos del tamaño de la población (ordenados de menor a mayor) con la que correspondería si fuese normal. A tal fin: a) se calcula por máxima verosimilitud la distribución normal acumulada (cdf) correspondiente a la de la población empíricamente observada (μ = 11,489; σ = 1,222); b) se obtiene por mínimos cuadrados el rango teórico que correspondería a dicha cdf (R^2 = 0,967); y c) finalmente se aplica el test W a ambas distribuciones (los rangos ordenados 1, 2 … 940 de forma creciente; y los rangos resultantes del análisis de regresión obtenidos por la cdf). Los resultados de la aplicación de esa última técnica permite, a diferencia del test KS, considerar la hipótesis de normalidad (p-value > alfa). Las dos distribuciones de rango observadas, puede afirmarse, corresponden a un mismo patrón.

Tabla n.1 Comparación de distribuciones por el método de Wilcoxon

V	210459
Esperanza	221135,000
Varianza (V)	69325822,500
p-valor (bilate	0,200
alfa	0,05

El p-valor exacto no ha podido ser calculado
El p-valor se ha calculado por aproximación

Sin embargo, ya como más adelante se discutirá, tampoco el test de Wilcoxon muestra resultados muy fiables, dada su naturaleza intrínsecamente ordinal.

Cabe concluir, por tanto, que para la muestra estudiada no existe una comprobación definitiva acerca de la normalidad de la distribución del logaritmo de la población de las áreas metropolitanas y micropolitanas USA[15]. *La aplicación, no obstante, del test W sugiere la posible existencia de una estructura oculta de la distribución del logaritmo de la población tendente a la normalidad*.

La tabla n. 2 resume los principales resultados de los diferentes tests ensayados para verificar la normalidad del logaritmo de la población, así como la hipótesis alternativa, relativa a la ley de Pareto.

Tabla n. 2: Tests de contraste

Hipótesis	**Log-Normal**	**Log-Normal**	**Log-Normal**	**Log-Log**	**Log-Log**	**Log-Normal**
Test	1 (W)	2 (W)	3 (W)	4 (W)	5 (W)	6 (W)
Contraste	O_I Pred Cdf[1]	O_I Pred cdf[1]	O_I_Pred cdf[2]	O_Pred pareto[3]	O_Pred pareto[4]	Prob. Acum.[5]
Tamaño	940	102	102	102	940	940
p-value	0,200	< 0,0001	0,526	0,276	<0,0001	0,730
alfa	0,05	0,05	0,05	0,05	0,05	0,05
Resultado	**Positivo**	**Negativo**	**Positivo**	**Positivo**	**Negativo**	**Positivo**

[1] Se trata del rango creciente predicho (para las 940 áreas Metro y Micropolitanas) a partir de la cdf. [2] Rango creciente predicho (para las 102 áreas Metropolitanas (más de 500.000 habitantes) a partir de la cdf. [3] Rango decreciente predicho a partir del modelo log-log (Pareto) para las 102 áreas Metropolitanas. [4] Rango decreciente predicho a partir del modelo log-log para el conjunto de las 940 áreas. [5] Probabilidad acumulada de la distribución empírica.

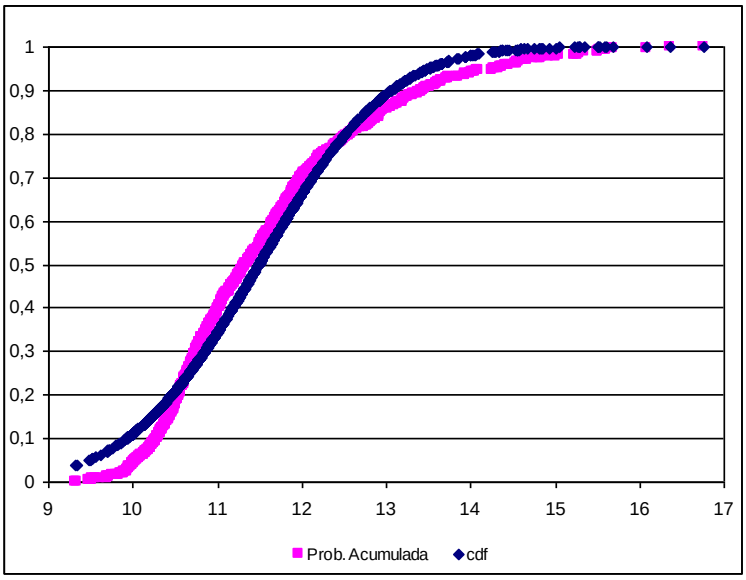
Como se puede observar de la tabla adjunta si bien el test 1 W permite hipotizar la estructura normal del logaritmo de la población *para el conjunto de la muestra* (las 940 áreas metro y micropolitanas), cuando se aplica al upper tail, las 102 áreas metropolitanas (2 W), dicha hipótesis no puede ser confirmada. En cambio si se confirma la distribución de Pareto (log-log, 4 W) para la cola alta, pero no así para el conjunto de la población (5 W). Nuestro trabajo permite contrastar, así, las conclusiones avanzadas por Malevergne, *et alt.* (2009) relativas a la corroboración de la distribución log-normal avanzada por Eeckhout (2004) para el conjunto de las ciudades USA, pero no para la upper tail, segmento en el que, en cambio, sería de aplicación la distribución log-log.

Los resultados parecen ser congruentes con la idea asentada de que las distribuciones de Pareto y lognormal exhiben diferencias cualitativas en sus respectivos upper tails. La densidad lognormal tiende a cero, en la cola alta, más rápidamente que cualquier densidad paretiana, lo que debería permitir distinguirlas claramente.

[15] La falta de demostración concluyente tal vez se deba a la estructura misma de la información empleada: las áreas metro y micropolitanas. Faltaría en esa información la población de los sistemas urbanos menores de 10.000 habitantes, los cuales son muy numerosos en los USA. Cabe recordar que, referida a *places* (>25.000 places en todo el país), Eeckhout (2004) demostró la validez de la distribución lognormal.

Ello no obstante, la hipótesis de normalidad del logaritmo de la población permanece con fuerza. El modelo 6 W, confirma la hipótesis nula relativa a la identidad de las distribuciones normal acumulada (cdf) y empírica acumulada (ver figura n. 7)[16].

Figura n. 7: Distribuciones normal y empírica acumuladas



Por su parte el test 3 W sugiere que, cuando el rango creciente se ajusta mediante un modelo de regresión con la cdf como variable dependiente tan sólo para el upper tail, parece confirmarse la identidad de ambas distribuciones. La cola alta de la distribución empírica podría denotar, por tanto, una distribución basada en la normalidad. Sin embargo dicho resultado no es congruente con el obtenido para el conjunto de la muestra.

[16] El modelo de regresión de ambas variables devuelve una R^2 = 979

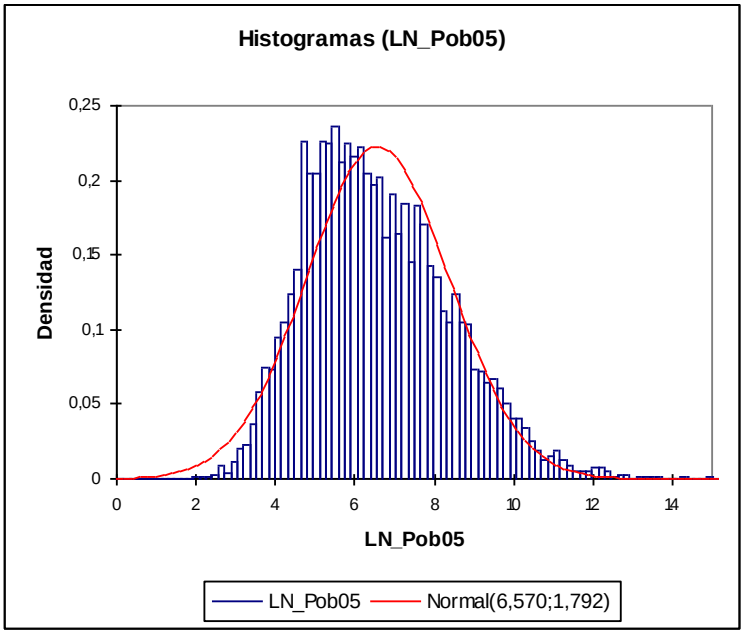
3.- La distribución de las ciudades en España

En segundo lugar replicaremos el estudio para los municipios y sistemas urbanos españoles. Los datos utilizados serán la población (referida a 2005) de los 8.109 municipios, así como la relativa (a 2009) de los 1.316 sistemas urbanos delimitados de acuerdo a la metodología que se expone más adelante. La razón de utilizar los sistemas urbanos, complementando a los municipios, consiste en contrastar si las estructuras reales (los *sistemas urbanos*) mejoran o no su rendimiento en relación a las estructuras de carácter administrativo, históricamente heredadas (los *municipios*).

3.1.- Los municipios españoles

El conjunto de pruebas clásicas relativas a comprobar la normalidad del logaritmo de la población vuelven a dar, como en el caso de las áreas metro y micropolitanas americanas, un resultado negativo, no permitiendo confirmar, prima facie, la hipótesis log-normal que parcialmente sugiere el histograma de la frecuencia de la población (figura n. 8)[17].

Figura n. 8: Histograma del LN de la población de los municipios españoles



A pesar del elevado poder explicativo (R^2 = 0,992) del modelo de regresión, con el rango creciente como variable dependiente y la cdf como variable independiente,

[17] De nuevo el histograma muestra una relativa falta de simetría en la distribución del logaritmo de la población. Los municipios con un logaritmo de su población situado entre 5 y 6 son significativamente más abundantes que los situados entre 6 y 7, haciendo dudar del carácter normal de la distribución, a pesar de su apariencia formal.

el test K-S no permite verificar la identidad entre la distribución de dicho rango y el valor predicho por el modelo de regresión resultante de la normal acumulada, tal como se deduce de la tabla n. 3 (test 1 KS). Identidad que, en cambio, si puede ser hipotizada en virtud del test W, de carácter ordinal (test 2 W).

Figura n. 9: Modelo basado en la cdf (totalidad de la muestra)

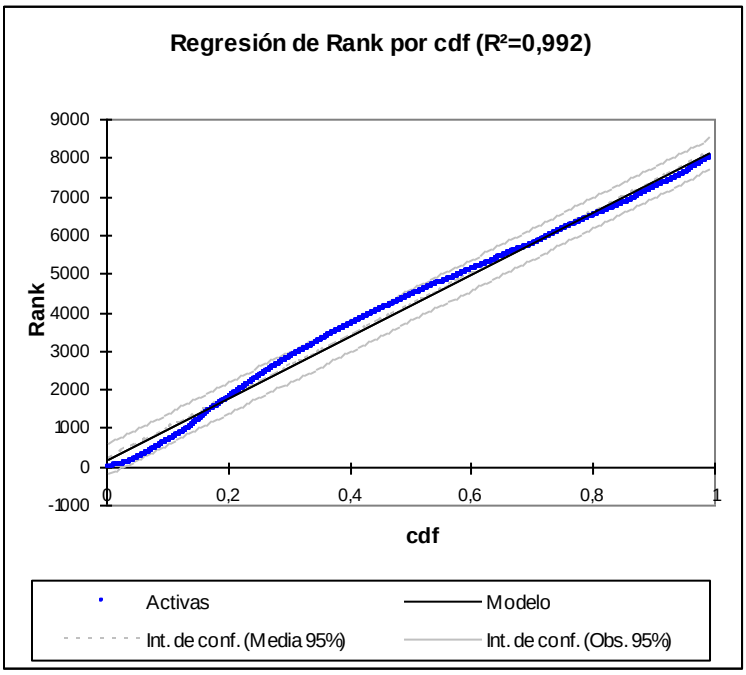


Tabla n.3: Tests de contraste para la población de los municipios (2005)

Hipótesis	Log-Normal	Log-Normal	Log-Log	Log-Log	Log-Normal	Log-Normal
Test	1 KS	2 W	3 KS	4 W	5 W	6 W
Contraste	O_I pred. cdf	O_I pred. cdf	O_pred. pareto	O_pred. pareto	O_I_N / cdf	O_pred. Pareto
Tamaño	8.109	8.109	58	58	8.109	58
p-value	< 0,0001	0,355	0,919	0,264	< 0,0001	0,927
alfa	0,05	0,05	0,05	0,05	0,05	0,05
Resultado	**Negativo**	**Positivo**	**Positivo**	**Positivo**	**Negativo**	**Positivo**

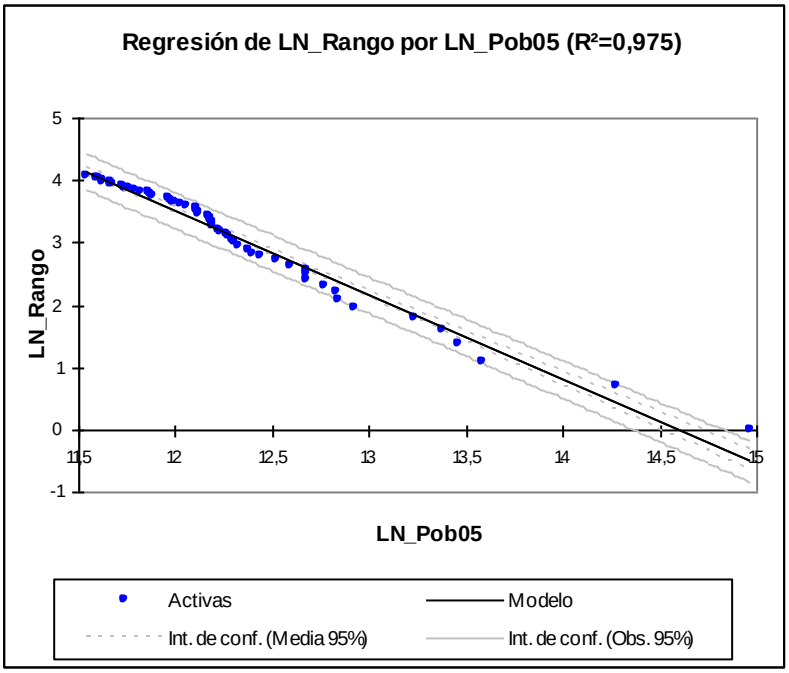
De nuevo, para el upper tail (en este caso los municipios con más de 100.000 habitantes), tanto el test de Kolmogorov-Smirnov (test 3 KS), como el de Wilcoxon (test 4 W) confirman la adecuación de la función log-log (que alcanza una bondad de ajuste del 0,975, figura n. 10). La Ley de Pareto queda, así, confirmada para los municipios más grandes españoles. No así la "ley" de Zipf, debido a que el coeficiente de regresión se muestra alejado de la hipótesis -1 (- 1,355)[18].

En cambio, y como ya sucedía en el caso de las áreas micro y metropolitanas USA, el modelo log-log fracasa cuando intenta explicar la distribución del conjunto

[18] Llama profundamente la atención la diferencia observada entre el coeficiente α en los modelos de las áreas urbanas USA y los municipios españoles. La mayor pendiente del modelo español sugiere una mayor macrocefalia en relación a la distribución de la población en los Estados Unidos.

de la muestra. La R^2, del 0,900, se aleja considerablemente del ajuste mostrado por el modelo log-normal, confirmándose, de esta manera, que la Ley de Pareto deja rápidamente de ser eficiente para la explicación de la distribución del conjunto de la estructura urbana, a diferencia de lo que sucede con la cdf resultante de la hipótesis de normalidad

Figura n. 10 Modelo log-log de los municipios españoles (100.000 habitantes)



Finalmente cabe indicar que si se replica un modelo de regresión del orden decreciente para los 58 municipios del upper tail con la cdf, se obtiene una bondad de ajuste relativamente elevada (R^2 = 0,952), si bien inferior a la alcanzada por la distribución log-log. El hecho, no obstante que tanto el test de Kolmogorov-Smirnov como el de Wilcoxon (6 W), sean positivos sugiere un comportamiento no deficiente de la cdf, si se ajusta un modelo específico para el upper tail.

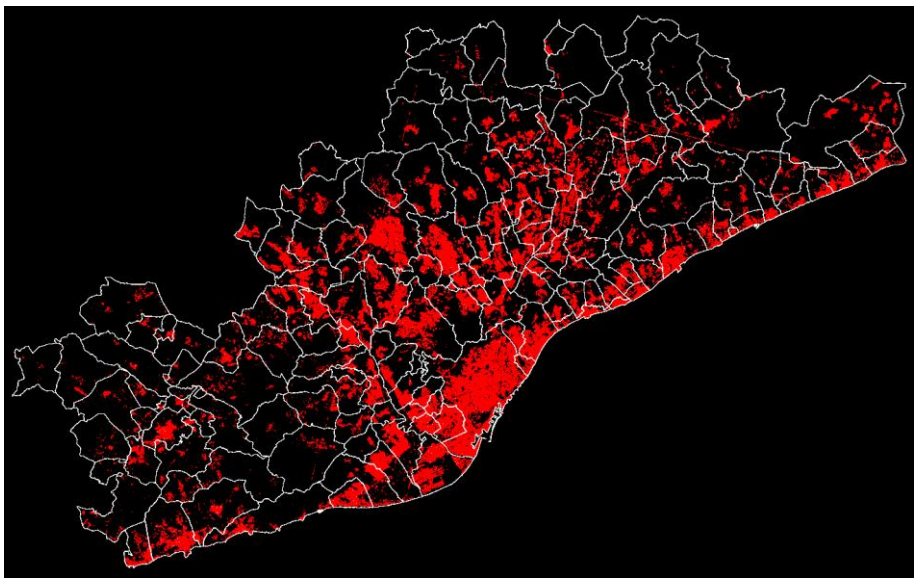
El ejercicio para los municipios españoles, por tanto, sugiere la validez de la hipótesis de normalidad del logaritmo de la población. Sin embargo el upper tail sigue mostrando una estructura que tiende a escaparse de la distribución log-normal, sugiriendo la existencia de elementos de singularidad en los municipios más grandes. Singularidad que sólo la hipótesis log-log de Pareto, parece capaz de explicar.

3.2. Los sistemas urbanos.

A fin de perfilar el análisis anterior con ciudades *reales*, y no con simples unidades administrativas (los municipios) se ha realizado una tarea encaminada a delimitar *sistemas urbanos*, más allá de los límites fijados administrativamente.

Por poner un solo ejemplo, decir que la “Barcelona” real corresponde al municipio de Barcelona (100 km^2, 1.593.080 habitantes en 2005) casi es tan absurdo como decir que “Londres” corresponde a la City of London, única entidad *local* británica que lleva el nombre de la capital del Reino Unido. La figura 11 muestra como la realidad urbana de Barcelona desborda, con mucho, su límite municipal (grafiado en blanco).

Figura n. 11: Suelo artificializado en la Región Metropolitana de Barcelona



Fuente: CPSV

La dificultad, sin embargo, consiste en obtener una metodología fiable de delimitación de las ciudades reales. No es objetivo de este trabajo profundizar en la discusión de formas alternativas de concreción de esas ciudades reales. Simplemente afirmar que no se propondrá que esas ciudades se identifiquen con las áreas metropolitanas. Las áreas metropolitanas se caracterizan por incorporar distintas realidades urbanas, físicamente continuas o no, caracterizadas por mantener entre ellas fuertes lazos de interacción. Sin embargo las metrópolis van más allá de la “ciudad”. Son auténticas “ciudades de ciudades”. Por ese motivo no creemos adecuada su utilización, del mismo modo que no lo son tampoco los municipios o entidades administrativas menores[19].

Para este trabajo se ha adoptado como metodología de delimitación, la utilizada por Roca *et alt.* (2009) relativa a la delimitación de *protosistemas* y *sistemas urbanos consolidados*, basada en la aplicación de la técnica del *valor de interacción* (Roca & Moix, 2005). Dicha metodología puede ser resumida por medio de los siguientes elementos fundamentales:

[19] Como por ejemplo los “places” utilizados en la reciente literatura dedicada a la discusión de las leyes de Pareto y Gibrat.

- Dada la matriz de flujos residencia/trabajo de base municipal (8.019 x 8.109) se calcula la matriz i/j de “valores de interacción” por medio de la ecuación:

$$VI_{ij} = \frac{F_{ij}^{\ 2}}{POR_i \cdot LTL_j} + \frac{F_{ji}^{\ 2}}{POR_j \cdot LTL_i}$$

  Donde VIij es el valor de interacción entre el municipio i y j, Fij y Fji, los flujos de i a j y de j a i, respectivamente, PORi y PORj la población ocupada residente de ambos municipios, y LTLi y LTLj los lugares de trabajo localizados en ambos municipios.

- Posteriormente se agregan los municipios en *protosistemas* en virtud de su máximo valor de interacción, de forma que esos protosistemas se cierran tan solo en el caso de que todos los municipios tengan su máximo valor de interacción con otro municipio de protosistema y que el conjunto sea físicamente continuo.

- Finalmente los protosistemas se consolidan en *sistemas urbanos* en caso de que la *autocontención*[20] sea igual o supere el 50%, puesto que los autores entienden que sólo puede llamarse “ciudad” aquellos sistemas urbanos capaces de retener al menos un 50% de la población ocupada residente[21].

Lo anterior permite la identificación de 1.531 *protosistemas*, de los cuales 218 no cumplen la condición de autocontención mínima (fijada en un 50%), lo que conduce a una delimitación definitiva de 1.316 protosistemas consolidados, que a los efectos del presente trabajo, serán considerados como *sistemas urbanos* reales. La figura n. 12 presenta los resultados de delimitación.

[20] Se entiende por autocontención el porcentaje de población ocupada residente que trabaja en el propio municipio (o protosistema).

[21] Ese 50% es la única condición impuesta a los sistemas urbanos. Por tanto no se impone ninguna condición *administrativa* del tipo de umbral mínimo de población o LTL.

Figura n. 12: Sistemas urbanos delimitados por medio del valor de interacción

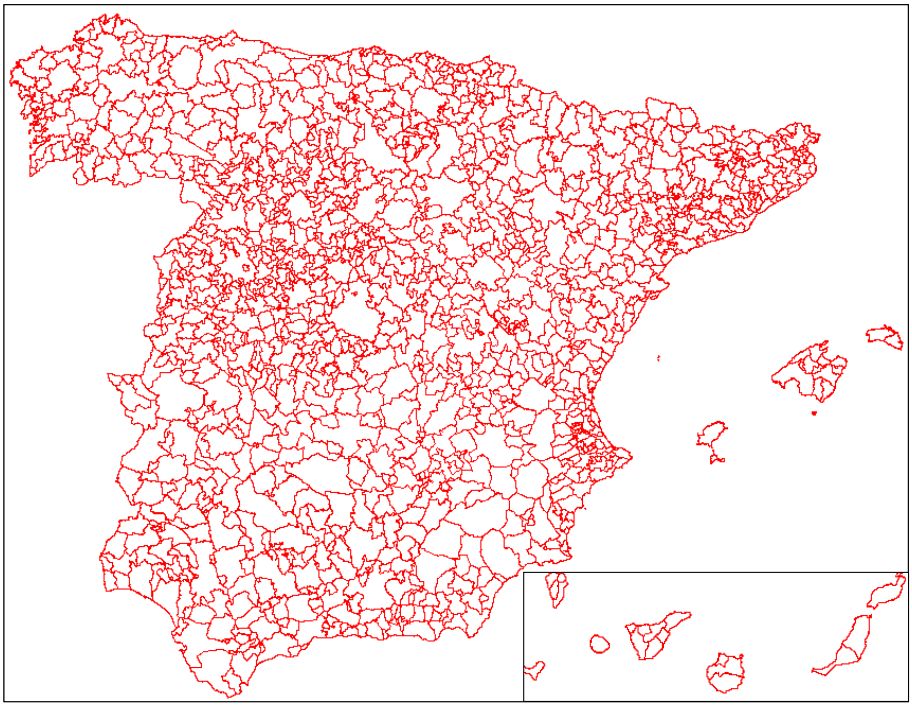
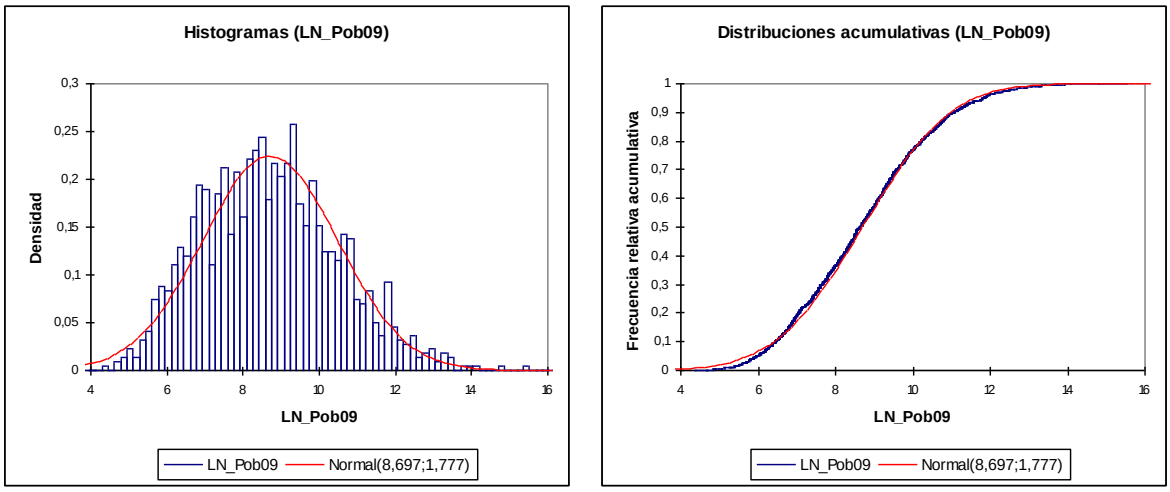


Figura n. 13: Histograma

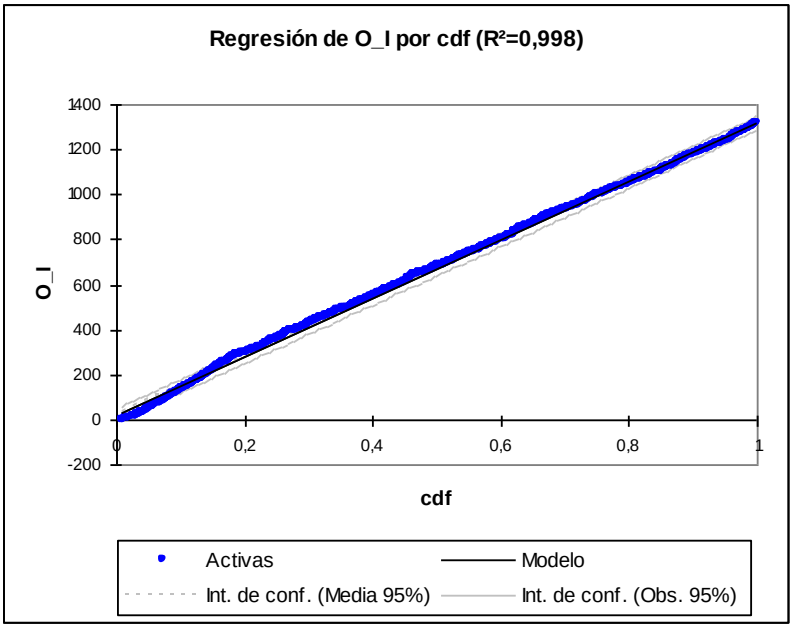


A pesar de que el histograma muestra una distribución que se acerca poderosamente a la normalidad, de nuevo el test estándar, de carácter paramétrico, no permite confirmar que la distribución del logaritmo de la población responda plenamente a una estructura normal, lo que obliga a buscar mecanismos alternativos de validación.

El ajuste de un modelo de regresión con el orden creciente (O_I) como variable dependiente y la cdf del logaritmo de la población, como variable independiente,

permite alcanzar una R^2 realmente espectacular: 0,9984, lo que permite hipotizar que efectivamente la población guarda observa una estructura log-normal.

Figura n. 14: Modelo de Regresión con la cdf como variable independiente



Sin embargo, y tal como se deduce de la tabla 4, el test de Wilcoxon (1 W) no permite, a diferencia de lo que sucedía en el caso de los municipios españoles y las áreas micro y metropolitanas USA, asegurar la identidad de ambas distribuciones (el orden creciente y el orden predicho resultante del modelo de regresión con la densidad normal acumulada calculada a partir del logaritmo de la población como variable explicativa). Dicho resultado negativo podría entenderse como una constatación de la no-normalidad del logaritmo de la población. Sin embargo, y en virtud de lo que se indica a continuación creemos no puede concluirse tal cosa, sino, al contrario, aparecen serias dudas de la validez del test de Wilcoxon para la corroboración de la identidad de distribuciones, dada su naturaleza intrínsecamente ordinal.

Tabla n.4: Tests de contraste para la población de los sistemas urbanos españoles (2009)

Hipótesis	**Log-Normal**	**Log-Normal**	**Log-Normal**	**Log-Normal**	**Log-Log**	**Log-Log**
Test	1 W	2 KS	3 KS	4 MW	5 KS	6 KS
Contraste	O_I pred. cdf	Distribución Normal	O_I pred. cdf	O_I pred. cdf	Orden pred. Pareto	O_I_N pred. cdf
Tamaño	1.316	1.316	1.316	1.316	88	88
p-value	0,005	0,130	0,478	0,999	0,987	0,986
alfa	0,05	0,05	0,05	0,05	0,05	0,05
Resultado	**Negativo**	**Positivo**	**Positivo**	**Positivo**	**Positivo**	**Positivo**

La aplicación del test de Kolmogorov-Smirnov, de carácter no paramétrico, para contrastar el ajuste a una distribución normal del logaritmo de la población (2 KS), obtiene un resultado positivo (p-value=0,1304), de forma similar al test KS dirigido

a comprobar la identidad entre la distribución correspondiente al orden creciente (O_I) y la resultante del modelo de regresión con la cdf como variable dependiente (3 KS). Como el p-valor calculado (0,478) es mayor que el nivel de significación alfa=0,05, se puede aceptar la hipótesis nula de identidad entre ambas distribuciones. La contradicción de resultados entre los tests W y KS pone en duda la metodología basada en la comparación simplemente ordinal. En el mismo sentido, la prueba de Mann-Whitney / prueba bilateral (ver modelo 4 MW) otorga resultados positivos a la comparación entre el orden creciente y la resultante del modelo de regresión con la cdf.

El conjunto de pruebas confirma, por tanto, la hipótesis log-normal de la población de los sistemas urbanos, con una contundencia antes no alcanzada para los municipios o las áreas metro y micropolitanas. *La estructura de la población de las ciudades reales parece ajustarse a una distribución log-normal*.

El contraste de la ley de Pareto se concentra en los 88 sistemas urbanos de más de 100.000 habitantes. El logaritmo de la población obtiene una R^2 = 0,991, con el logaritmo del rango (orden decreciente), confirmando una vez más el excelente comportamiento del modelo log-log en el upper tail. Por su parte el test no paramétrico KS (5 KS), como también el de Wilcoxon, permite confirmar la identidad entre la distribución del logaritmo del rango y el valor predicho del mismo resultante del modelo de regresión con el logaritmo de la población como variable independiente.

Sin embargo el modelo del upper tail, con la cdf como variable independiente y el orden inverso normalizado (O_I_N), otorga la sorpresa de alcanzar un nivel de explicación aún más elevado (R^2 = 0,993), así como la confirmación, por los tests de KS (6 KS) y W, de correspondencia entre ambas distribuciones.

4.- Conclusiones

La realización de los estudios precedentes permiten, para las áreas micro y metropolitanas USA, obtener las siguientes conclusiones:

1. La estructura de la población del conjunto de las áreas urbanas USA de más de 10.000 habitantes parece corresponder a una distribución log-normal, tal como ha propuesto Eeckhout (2004) para el conjunto de las ciudades USA.
2. Dicha conclusión no parece poder elevarse al upper tail, segmento en el que, en cambio, sería de aplicación la distribución log-log, tal como ha sugerido Malevergne, *et alt.* (2009).

Idénticos resultados parecen observarse del estudio de la estructura de la población de los municipios españoles. De nuevo para el conjunto de la muestra el modelo log-normal viene a ser confirmado. Sin embargo en el upper tail continúa dando muestras de debilidad, mostrando una clara supremacía el modelo log-log.

Sin embargo las conclusiones anteriores no pueden ser dadas como definitivas, dada la naturaleza de los datos estudiados en las dos series anteriores. En ninguno de ellos se han considerado las *ciudades reales*. En el caso de los USA, puesto que si bien las áreas metro y micropolitanas podrían obedecer a ciudades reales, no han podido ser considerados los sistemas urbanos de menos de 10.000 habitantes, muy abundantes fuera de los entornos metropolitanos. En el caso de España, puesto que los municipios responden a entidades administrativas, pero no representan un reflejo fiel de la realidad de la estructura urbana del país.

De forma alternativa se ha contrastado la metodología desarrollada por Roca *et alt.* (2009), relativa a la delimitación de sistemas urbanos por medio del sistema del valor de interacción. Y el resultado parece confirmar la hipótesis de que *cuando nos encontramos con ciudades reales, la distribución de la población responde de forma más acentuada a una estructura log-normal*. La mejora de prácticamente todos los indicadores empleados (ver tabla n. 5) sugiere, de forma contundente, la necesidad de mejorar los trabajos empíricos utilizando no sólo la totalidad del universo urbano, sino, además, la consideración de sistemas urbanos reales.

Tabla n. 5

Muestra	R^2 OIN-cdf[1]	KS Norm[2]	KS OIN-cdf[3]	KS OI-pred_cdf[4]	MW OIN-cdf[5]	MW OI-pred_cdf[6]
Municipios	0,992	< 0,0001	< 0,0001	< 0,0001	0,00244041	0,99116671
Sistemas Urbanos	0,998	0,13038	0,47755912	0,47774589	0,00462283	0,99872660

[1] Modelo de regresión entre el orden inverso normalizado (OIN) y la densidad normal acumulada (cdf) del logaritmo de la población. [2] Contraste de normalidad de Kolmogorv-Smirnov. [3] Comparación de identidad de las distribuciones del OIN y la cdf por medio del test de Kolmogorov-Smirnov (KS). [4] Comparación por medio del test KS de la identidad de las distribuciones del rango creciente (OI) y la predicción del mismo a partir de un modelo de regresión con la cdf como variable independiente. [5] Comprobación, por el test Mann-Whitney (MW), de la identidad entre el OIN y la cdf. [6] Comprobación por el test MW de la identidad entre el rango creciente y y la predicción del mismo a partir de un modelo de regresión con la cdf como variable independiente.

Bibliografía

Alperovich, G., (1993): "An Explanatory Model of the City-Size Distribution: Evidence from Cross-country Data", *Urban Studies*, 30, 1591–1601.

Berry, B.J.L. (1961): "City Size Distributions and Economic Development", *Economic Development and Cultural Change,* 9, 593–587.

Black, D. & Henderson, J.V. (2003): "Urban Evolution in the USA", *Journal of Economic Geography*, 3, 343-372.

Berry, B.J.L. & Horton, F.E. (1970): *Geographic perspectives on urban systems*, Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

Carroll, G. (1982): "National city size distributions: what do we know alter 67 years of research?", *Progress in Human Geography*, 6, 1-43.

Cheshire, P.(1999): "Trends in Sizes and Structure of Urban Areas", en *Handbook of Regional and Urban Economics* (P. Cheshire, and E. S. Mills, eds.), Elsevier Science, B. V., Amsterdam.

Dobkins, L.H. & Ioannides, Y.M. (2000): "Dynamic Evolution of the US City Size Distribution", in *The Economics of Cities* (J.-F. Thisse and J.-M. Huriot, eds.), Cambridge University Press, Cambridge (2000).

Eeckhout, J. (2004): "Gibrat's law for (all) cities", *American Economic Review* 94, 1429-1451.

Eeckhout, J. (2009): "Gibrat's law for (all) cities: Reply", *American Economic Review*, 99:4, 1676–1683.

Krugman, P. (1999): "El tamaño de las ciudades" en *The Spatial Economy* (Fujita, M., Krugman, P. & Venables, A.J. eds), Massachusetts Institute of Technology.

Gabaix, X. (1999): "Zipf's Law for Cities: An Explanation", *Quarterly Journal of Economics*, CXIV, 739–767.

Gibrat, R. (1931): *Les inégalités économiques*. París, Librairie du Recueil Sirey.

González-Val, R., Lanaspa, L., Sanz, F. (2008): *Nueva Evidencia sobre la Ley de Gibrat en Ciudades*. Universidad de Zaragoza.

Krugman, P.R. (1996): *The Self-Organizing Economy*, Blackwell Publishers, Oxford.

Lasuén, J. R., Lorca, A. y Oria, J. (1967): "City-Size Distributions and Economic Growth", *Ekistics*, vol. 24, págs. 221-226.

Lanaspa, L., Perdiguero, A.M., Sanz, F. (2004): "La distribución Del tamaño de las ciudades en España", *Revista de Economía Aplicada*, 34, vol. XII, 5-16.

Levy, M. (2009): "Gibrat's law for (All) cities, A Comment", *American Economic Review*, forthcoming.

Malevergne, Y., Pisarenko, V. & Sornette, D. (2009): "Gibrat's Law for Cities: Uniformly Most Powerful Unbiased Test of the Pareto Against the Lognormal". *Swiss Finance Institute Research Paper* N. 09-40, September 2009. http://ssrn.com/abstract=1479481.

Pareto, V. (1896): *Cours d'Economie Politique*. Geneva, Droz.

Parr, J. (1985): "A Note on the Size Distribution of Cities over Time", *Journal of Urban Economics*, 18, 199-212.

Roca, J., Marmolejo, C. & Moix, M. (2009): "Urban Structure and Polycentrism: Towards a Redefinition of the Sub-centre Concept", *Urban Studies*, 46, 2841–2868.

Roca, J. & Moix, M. (2005): "The interaction value: its scope and limits as an instrument for delimiting urban systems", *Regional Studies* 39, 357-373

Rosen, K.T. & Resnick, M. (1980), "The size distribution of cities: an examination of the Pareto law and primacy", *Journal of Urban Economics*, 8:165-186.

Suárez-Villa, L. (1988): "Metropolitan Evolution, Sectoral Economic Change, and the City Size Distribution", *Urban Studies*, 25, 1-20.

Zipf, G. (1949): *Human Behavior and the Principle of Least Effort*, New York, Addison-Wesley.